eurostat: Eurostat Open Data R Tools

Leo Lahti, Janne Huovari, Markus Kainu, Przemyslaw Biecek

Governmental institutions are increasingly opening up their data resources for the public as open data. This is providing novel opportunities for research and citizen science, but efficient tools to access and analyze these data sets are needed to realize the full potential of the new information resources. We introduce the eurostat R package that provides a suite of tools to access open data from Eurostat, including functions to search, download, and manipulate Eurostat data in an automated and reproducible manner. The online documentation provides detailed examples on how to access, summarize and visualize these spatio-temporal data sets. The package expands previous related work and has been extensively tested by the user community. This contributes to the growing ecosystem of R packages that provide algorithmic tools for reproducible computational research in social science and humanities.

Eurostat, the statistical office of the European Union, provides a rich collection of demographic and economic data through its open data service, which currently includes over 8800 data sets on European demography, economics, health, infrastructure, traffic and other topics. In many cases the statistics are available with great geographical resolution and including time series spanning over several years or decades.

The availability of tools to access and analyse data collections from the public domain can greatly benefit reproducible research Gandrud13, Boettiger2015. When the data resources and analysis algorithms are openly available, the complete analytical workflow spanning from raw data to the final publication can be made fully automated and transparent. Standardization of common data analysis tasks via dedicated software packages can help to automate the analysis workflow, greatly facilitating reproducibility and code sharing, and making the data analysis more efficient. The algorithms need to be customized to specific data sources, however, to accommodate variations in raw data formats, access details, and typical use cases so that the end user can avoid repetitive standard programming tasks and spend more time on the actual research tasks. A number of packages to access specific data sources from governmental and other institutions have been consequently designed to meet these demands and to access open data from the Food and Agricultural Organization (FAO) of the United Nations (FAOSTAT; FAOSTAT), World Bank (WDI; WDI), national statistics authorities (pxweb; pxweb), Open Street Map (osmar; osmar) and many other sources.

A dedicated R package for eurostat open data has been missing, however. We introduce the eurostat R package to fill this gap. The package facilitates automated access to open data from Eurostathttp://ec.europa.eu/eurostat/data/database This brings together our earlier efforts with the statfi statfi and smarterpoland smarterpoland packages. Compared to this earlier work, we have now combined the relevant parts of these two packages and implemented an expanded set of tools with a specific focus on the Eurostat data collection. The first CRAN release of the package was in 2014. Since then it has been actively developed by several contributors and based on community feedback in Github. We are now reporting the first mature version of the package that has been improved and tested by multiple users. The package and its predecessors have been applied in several case studies by us and othersSee e.g. http://blog.revolutionanalytics.com/2015/04/financial-times-tracks-unemployment-with-r.html

Related work includes the datamart datamart and the quandl quandl R packages that provide generic tools that can be used to access certain versions of Eurostat data. In contrast to these generic database packages, our eurostat package provides functionality that is particularly tailored for the Eurostat open data service. The development version of another related R package reurostathttps://github.com/Tungurahua/reurostatoes not seem to be actively maintained at the moment. Moreover, our eurostat package depends, imports or suggests the following external R packages: devtools devtools, dplyr dplyr, knitr knitr, ggplot2 ggplot2, mapproj mapproj, plotrix plotrix, reshape2 reshape2, rmarkdown rmarkdown, stringi stringi, testthat testthat, and tidyr tidyr. The eurostat R package is part of rOpenGov collection Lahti13icml that provides reproducible research tools for computational social science and digital humanities.

In summary, the eurostat package provides custom tools to search, retrieve, modify and visualize data from the Eurostat open data service. The package supports key features such as data cache, date formatting, and tidy data principles wickham2014 using the the tidyr R package tidyr. Here, we provide an overview

ostat/data/database .

loyment-with-r.html .

ngurahua/reurostat d

of the core functionality in the current CRAN release version (1.2.1). For further examples, see the package vignettehttps://github.com/rOpenGov/eurostat/

Search and download commands

To install and load the CRAN release version, just type in R:

example ¿ install.packages("eurostat") ¿ library("eurostat")

OpenGov/eurostat/ .